DOES DATA MINING IMPROVE BUSINESS FORECASTING?

June 13, 1998

David Chereb, Ph.D.

Prepared for:
THE 18TH INTERNATIONAL SYMPOSIUM ON FORECASTING

Edinburgh, Scotland

INTRODUCTION

The purpose of this article is to show why naïve data mining will produce misguided business forecasts. Additionally, this article shows how enhanced data mining techniques add value to business decision-making.

In order to show how enhanced data mining can add value to decision-making, we must first show how realistic models work. Many times I've fallen into the trap of using a model structure because I like it (or just learned it), instead of fitting the model to the problem at hand. For today's data miners, this is a common danger. To prevent 're-inventing the wheel' this article shows how to make a 'real world' model that is used and relied upon.

The next section shows the characteristics of both good and bad business forecasting models. For non-business forecasting models (climate, chemistry, etc.) the details are different. This is our first clue that forecasting models are 'domain dependent' (i.e. a model works best within a particular, narrow environment only).

The section on data mining strips away the 'hype' and shows why it is very useful 'sometimes'. This section also shows how to avoid all the mistakes that naïve forecasters have been making for the past 100 years.

The section on robust models gets to the heart of how to build and use models that work. An example of all these characteristics, good and bad, are presented in the housing forecast section. The data set consists of annual, single family housing starts for each U.S. County over a ten year period (about 30,000 data points). The models attempt to forecast future annual housing starts by county.

The article finishes by summarizing the forecasting challenge into five rules of thumb.

WHEN DOES DATA MINING MAKE THINGS WORSE?

Data mining makes business forecasting worse when the model framework does not incorporate economic theory and complexity theory. Two of the most important characteristics of business dynamics are substitutions and adaptation. Neither of these characteristics is automatically accounted for in data mining algorithms. In addition, one of the most typical characteristics of today's business environment is structural change. Forecasting structural change is a complex, difficult task that is not easily codified. Very few of today's data mining algorithms do decent job of capturing structural change.

Those who use data mining without encapsulating it in the proper modeling framework will miss important turning points and mislead users. At best, the results may come close to a properly designed model, and at worst the results will lead to disastrous decisions. Any reasonable business forecast, say an annual budget model, has to account for a changing environment. Incorporating the changing environment is more important than the data set.

For instance, if a strong increase in sales is projected, but the model ignores the fact that a competitor has just dropped the price of a competing product, the forecast will be wrong. It doesn't matter if the data set has 10 million records. What matters is that a changing environment is accounted for.

BUSINESS FORECASTING CHARACTERISTICS

Good Forecast vs Good Model

From a technical perspective we want a model that is consistent, solvable and statistically valid; but none of this matters in a business forecasting model. The only thing that matters is if it does a good job of forecasting. True, a model that forecasts well and meets all of the technical criteria is preferred. But we've never seen a situation where technical merit counts more than practical results. And this is as it should be in business forecasts. Accurate business forecasting models are also much harder to produce than technically sophisticated models. Our goal is to show how to construct good business forecasting models.

DATA MINING CHARACTERISTICS

For the purposes of this article, data mining is defined as the semi-automatic search for hidden relationships in large databases.

Our focus is using data mining for business forecasting. Within this environment, raw data mining, even with all the new tools and rising popularity, remains suspect. While data mining is a very useful tool, it uses only the information in the data set to solve a problem. In essence data mining uses whatever data resides in one or more databases. That is the Achilles heel for any model because a tremendous amount of information is not in the database and therefore gets ignored. Seasoned business decision-makers use well tested 'cause-n-effect' models in their toolkit.

Database Focus

Databases are good places to store a lot of data. If spreadsheets could easily handle a few billion rows and a few million columns, we might do most of our data mining in a spreadsheet. But for now, databases are more efficient at handling millions and millions of bits of data. In addition, databases are becoming smarter with the growing use of data warehouses and OLAP. These tools make analyzing the data, instead of just storing the data, much easier.

The last big reason that data mining is so database oriented is that the big software companies are trying to sell more database engines. They are beginning to put a lot of advertising dollars into data analysis and business intelligence analysis, via databases.

 dc@davidcherebgroup.com

Automatic Searches

Most data mining products allow for semi-automatic searching and analyzing. This is a big convenience and most welcome when your data set is in the gigabyte range. This is also more valuable the less you know about the domain. When I know almost nothing about the domain, an automatic search will quickly educate me. The more I know, the less I need an automatic search capability. When I know about the domain, I mainly want to test specific actions, such as how much advertising I need do to increase sales next quarter by 15%, when I've heard that my closest competitor will raise prices by 5%.

Few Constraints

Data mining algorithms apply almost no constraints on model behavior (especially neural networks). This is touted as a strength, but in business forecasting it is more often a problem for two reasons. One, the model will reinvent the wheel by 'discovering' things that are obvious; and two, the model will produce results that violate theory (such as the laws of supply and demand).

For instance, some models try to forecast the impact of advertising on product sales by only including data on sales and advertising. They may have sales data by store and advertising by week, making it seem as if there are thousands of data points. In reality there are only two data points: sales and advertising. No matter how many data mining algorithms are applied to the data (neural nets, genetic, k nearest neighbor, regression, etc.) they will not provide very useful forecasts.

In the above case it may seem obvious that the models are not very good because they don't include competitors actions, the impact of substitute products, and a host of things any marketing manager must evaluate.

Models should have constraints. Constraints are a type of information and enhance the robustness of the model.

Why Now?

Data mining is gaining favor now due to more powerful computer hardware and software, at lower prices. Everything else, the volumes of data, the time constraints, etc. are secondary. If data mining tools cost a minimum of $225 million, the market would be very limited. As the total package price of data mining solutions drops, the market is expanding. And as before, Microsoft Corporation will expand the market with low priced products. The combination of more field experience with data mining, and low prices, means that data mining usage will grow well over 50% per year for the next five years.

Advantages and Disadvantages

The advantages of data mining models are that they can handle large volumes of raw data and can be easily updated with new information.

The disadvantages of these same models are that they do not incorporate economic theory and do not easily forecast structural change.

 dc@davidcherebgroup.com

ROBUST MODELS

As shown in prior sections, data mining can make business forecasting worse off. The solution is to add economic theory and complexity theory into the model. The result is a robust business forecasting model that is used and useful.

What Data Mining Business Forecasting Models Need

- **Economic Theory**
- **Complexity Theory**

Economic Theory

Applying economic theory constrains the model to behave according to proven laws. For example, the law of demand says that when the price of a product decreases, with other things held equal, the demand for the product will increase. This translates into forcing the model to yield a negative coefficient for a unit price variable. The data set, however, and data mining results will permit a positive coefficient, depending on the mix of data used. But a positive coefficient violates an economic law and will lead to ridiculous business decisions. Other economic laws that constrain and improve business forecasting models are: cross-elasticities of demand and time dependent elasticities.

Complexity Paradigm

The Complexity Paradigm, as used here, means the use of feedback channels and adaptive behavior. The feedback mechanism allows the model to react over time to either prior or current changes. It adds an extra level of realism to the forecast.

In addition the Complexity Paradigm permits the model to adapt over time. This means the reaction to say an increase in advertising, will change during the forecast period. This is very different from most models. What this is saying is that the forecasting coefficients are time and situation dependent, and change as the forecast unfolds.

These two additions to a data mining model add robustness and realism.

 dc@davidcherebgroup.com

Bayesian Rules

As we know, the formal method for incorporating additional information is through Bayes Rule. Within the context of business forecasting, this additional information can be added before or after the model is run.

While some may denigrate using 'add factors', 'fudge factors', or intuition, not to use these things is to make an unrealistic assumption that the model contains everything relevant to the forecast. This is crazy. There are trillions of factors and combinations that can impact the forecast. We ignore most of them simply because our brains, tools and models can't handle it. Ignoring almost everything is a Bayesian input. Thus we're all Bayesians. The 'invisible hand' of market dynamics allows us to ignore most variables because we know that we do not have to produce and solve equations for every transaction. The 'invisible hand' of the market insures that the market tends toward equilibrium (especially if we allow enough time for substitutes). Adding this economic law saves about 99% of the work in building a business forecasting model.

The next few paragraphs show how Bayesian forecasting models improve business forecasts. We will use two notations, one mathematical and the other symbolic.

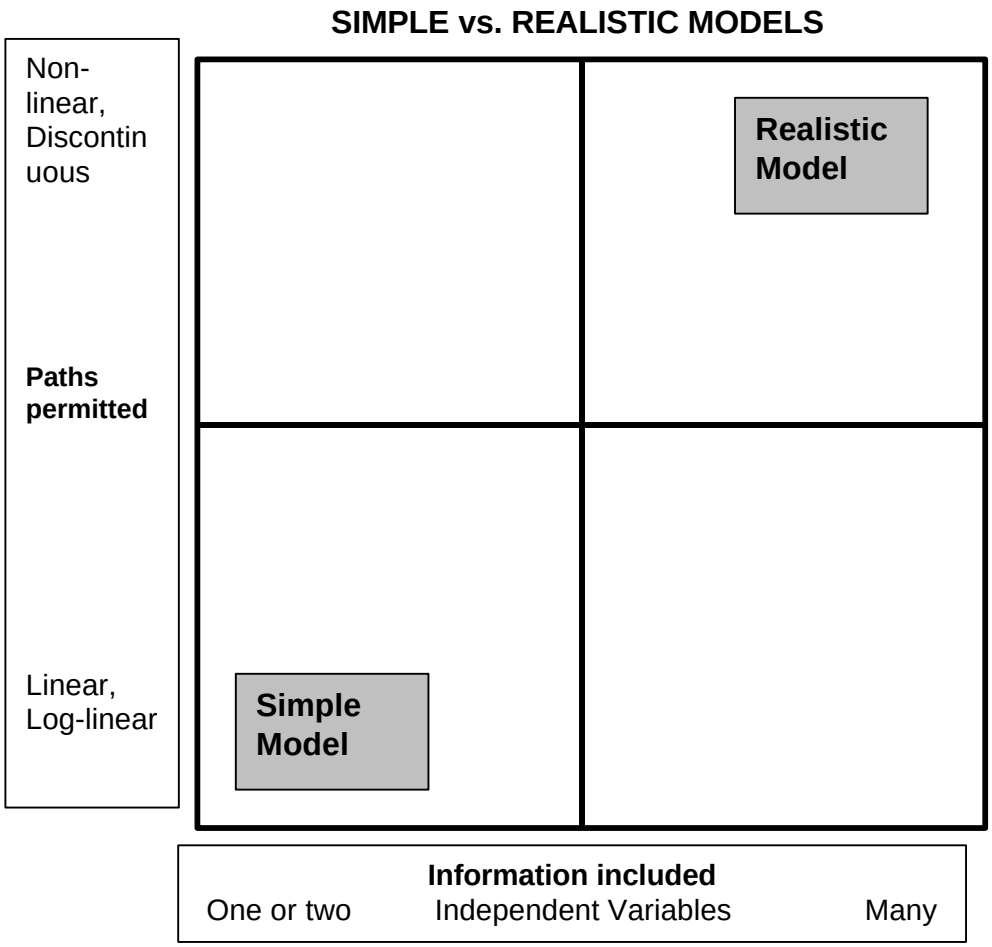
English	Math	Pseudo math
Given additional information, the revised forecast is	P(Y/X)=P(YandX) / P(X) = P(Y)P(X/Y) / P(X)	P(Y given evidence X)= P(Y alone)P(information that X adds)
Does the new information make any difference?	Ho: = P(Y/X) = P(Y) H1= P(Y/X)<>P(Y) If -2 < z < 2 then reject Ho	X must add relevant information, before the forecast of Y is improved.
If the new information costs money, then it has to improve the forecast by more than it costs.	EVI > Cost(I) Where EVI = w1EV1 + w2EV2+ . . And EVI=Expected Value of Information W1=weight (0-1.0) of information 1 W2= weight of information 2, etc.	The potential payoff with the new information > cost of information.

The above shows that information, whether in a database or from a manager's judgment, improves the forecast when it yields more in accuracy than it costs to acquire.

Forecasting Levels: Simple to Realistic

We classify business forecasting models using two criteria: one, what kind of feedback paths do they permit (Complexity Theory) and two, how much information is included in the forecast (Economic Theory).

The next chart shows our assessment of models, from simplistic too realistic.



In simple models the forecast can only be linear or log-linear. In a realistic model the forecast path can non-linear and discontinuous (i.e. able to show an abrupt change due to legal or technological events). Separately, but related, a simple model includes very little independent information. A simple model usually contains only past sales, but not advertising, competitor prices, new competing products, cannibalization of your own products, etc. There is almost no chance that a simple model will forecast correctly.

Two examples may help to illuminate this chart. Let one business forecasting model include daily sales by product, by store location, for the last five years. For 25 products and 200 store locations, this adds up to 1.85 million data points (365x5x25x200). Suppose the model used to forecast next year's sales by product, by store location uses data mining or some form of Box-Jenkins. Suppose further that the 'driver variables' are past sales by product, by store. To most business forecasters, and me this is a very simple model. The next table shows simple vs realistic for this problem.

 dc@davidcherebgroup.com

SIMPLE vs REALISTIC SALES FORECASTING MODEL EXAMPLE

	SIMPLE	REALISTIC	COMMENTS
Data points	1.850 million	1.851 million	In actual data points there is almost no difference. The extra information for the realistic model may consist of one extra data point, such as a management decision to add 2% to the model's results to account for improved quality.
Allow discontinuous points in forecast?	No	Yes	The realistic model will allow events such as legal or technical changes to cause a discontinuous path.
Incorporate competitor actions?	No	Yes	
Driver Variables	1	9	Driver variables are information. Past sales are only one piece of independent information, even if you have millions of data points.
- Past sales	X	X	
- Local employment changes		X	
- Average competitor prices?		X	
- Average price of substitute products?		X	
- Change in advertising $?		X	
- Number of Judgment variables?	0	5	▪ Such as: quality of management ▪ Customer satisfaction rating. ▪ New competing products to be introduced.

 dc@davidcherebgroup.com

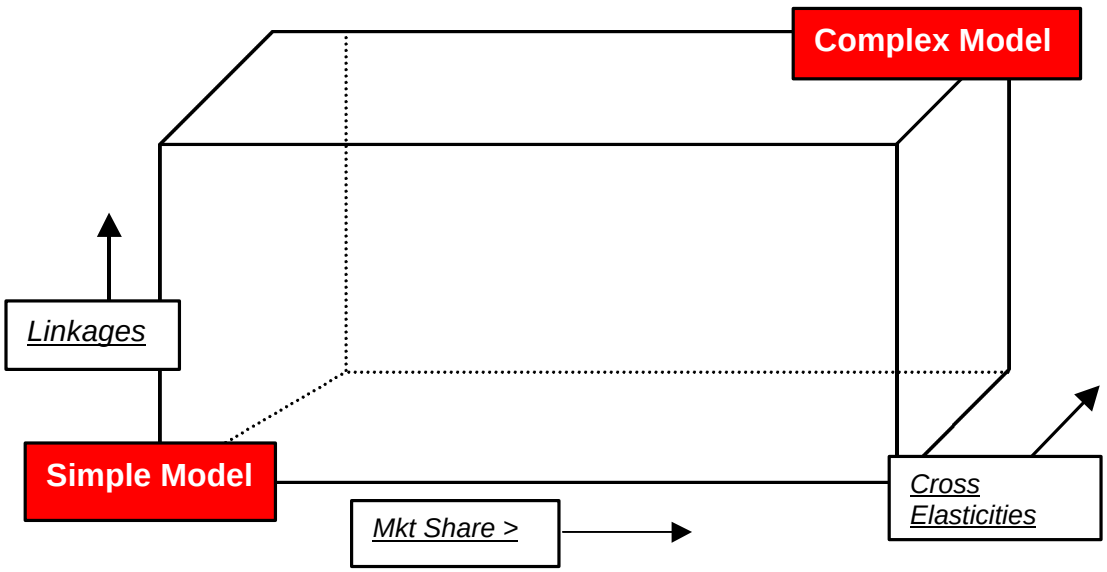
Complexity

One model does not fit all. Every business model is domain and path dependent. That is, a sales model for a mom-n'-pop shoe store should be different than a foreign currency forecasting model. Beyond this obvious statement comes the task of trying to fit model types into a reasonable framework. One attempt is shown next.

For an internal company forecasting environment, the next chart shows how to determine what type of model is appropriate. Model design is more important than the data. The design determines the basic characteristics of the model (much like classifying whether an entity is plant or animal).

To a beginner, all models are probably complex and to a seasoned forecaster, all models are simple. Whether the model forecasts correctly is another matter. *To us a model should be simple (i.e. small with few feedback loops) if it fits in the bottom, front, left of the cube. This is where there are weak linkages to external events, the product market share is small, and there are few substitutes*

SYNCHRONIZING MODEL COMPLEXITY with FORECASTING DOMAIN



High Linkages = Heavily influenced by external conditions.
High Market Share = Dominant company in that industry.
High Cross Elasticities = Easy for users to use substitute products.

	LINKAGES	MKT SHARE	CROSS-ELASTICITIES
Internal *Decisions Drive Company*	***Low***	***High***	***Low***
External *Conditions Drive Company*	***High***	***Low***	***High***

The opposite applies when the opposite conditions exist. When external events frequently have a strong impact on internal sales; when your company has a very high market share; and when there are close substitutes (i.e. high cross-elasticities), the forecasting model should be complex. As Einstein said, “you should try to make things as simple as possible, but not too simple”.

Most of the time forecasters error with models that are too simple because they let data constraints dictate model design. At times, a reasoned business judgment answer is more accurate than a data intensive model because the decision-maker is using a better model design. In fact it can be stated that once the problem can be fully ‘codified’ it is a simple problem. Understanding how an object reacts with gravity is simple since the equation is simple, universal and unchanging. Understanding how the stock market will behave over the next 18 months is not simple, and not easily codified.

Valuable Models

In business forecasting a valuable model is one that forecasts correctly and is used. Whether the model is elegant is secondary. Most seasoned business forecasters modify the model results with a little judgment when company profits are on the line.

REGIONAL HOUSING FORECASTING MODEL

Available from the author. Please send e-mail requesting model details.

SUMMARY

This article has provided a framework for producing realistic business forecasting models. The addition of data mining tools into business forecasts is a small step forward. The reason data mining is not a big step forward is that model design is still the most important ingredient for any business forecast. Below are five rules of thumb to insure your model is realistic and useful. Each of these has been discussed in the prior paragraphs. Together these rules of thumb will keep your model within reasonable bounds, and prevent many wasted labor hours.

FIVE RULES OF THUMB FOR BUSINESS FORECASTING MODELS

1. **Model Design is more important than data crunching.**
2. **Constraints and economic theory add to a model's robustness.**
3. **We're all bayesians (because every model includes judgments).**
4. **Models are domain and path dependent.**
5. **Complex environments require complex models.**

The final point is that good business forecasting is hard. I've been doing it for a long time, in many different environments, and it remains a challenge. The successes have come after many bruises and many dead-end paths. Today more than ever, the complex, rapid pace of business, requires models to be well grounded in economic and complexity theory.

David Chereb, Ph.D.
949-458-7794
Fax 458-9084
dc@davidcherebgroup.com
www.davidcherebgroup.com

ABSTRACT

DOES DATA MINING IMPROVE BUSINESS FORECASTING?
Author: David Chereb, Ph.D.

Keywords: data mining, bayesian, feedback system, structural change, business forecasting

Under some circumstances data mining results reduce the accuracy of business forecasts. The conditions under which this occurs are common in business projections that must include structural change. This data-mining anomaly can be prevented with proper design techniques. The circumstances under which this anomaly occurs are presented in this paper, along with the techniques needed to prevent this undesired result.

The paper presents examples and practical guidelines for improving forecasting results through data mining. The analysis centers on Bayesian techniques for incorporating apriori knowledge into the data set. This preconditioning of the data set reduces the negative impact of naïve assumptions. In addition the guidelines show how to incorporate structure change through a feedback mechanism. The net result is a robust, adaptive forecasting system. Examples from demographics are used to forecast housing starts by region.

David Chereb, Ph.D.
949-458-7794
Fax 458-9084
dc@davidcherebgroup.com
www.davidcherebgroup.com